Better Decision Making for New York City Taxi Drivers

Kefu Zhu, Chunlei Zhou

Problem Statement

As widely recognized symbols of the New York City, yellow and green taxis have a profound history since 1890s. Taxicabs provide both tons of employment positions and convenient commute options for the people in New York City under the supervision of Taxi and Limousine Commission (TLC). As TLC made the taxi trip records available to the public, a wide range of analysis has been done based on the data. Previous studies using the New York City Taxi data including how the trend changes among different types of taxis across time and area [1], what are the factors that influences the ride duration [2], what is the average speed of a trip at a particular hour of the day [3], where are the hot pick-up/drop-off areas in the New York City [4] and many other interesting topics [5].

In this project, we will explore the same data source from TLC[1] and study how can a taxi driver make better decisions, as well as the customer profile of taxi passengers within the New York City. We have several thoughts at this point and may make some adjustments given the limited time budget of this project. Firstly, we will investigate where should the drivers go for pickups if they want higher tip earnings for different types of taxis, which will be measured in both tip amount and tip as a percentage of total fare. Secondly, we will explore the effect of weather and time on the daily income of drivers respectively. Lastly, we would like to discover the properties of customers of different categories for various kinds of taxi.

[1] TLC Trip Record Data: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Data Acquisition

The data will be to programmatically download and loaded into python from the TLC website, which provides different datafile in CSV format every month from January 2009 to June 2019. We plan to use the 2018 data for different types of taxi and analyze our objectives across different months. Take the January 2018 yellow taxi data as an example, it has 8,759,874 records with 17 attributes. The detailed dictionary of attributes is also provided on the TLC website[2].

Algorithms

We may implement one or more of the following algorithms to explore the customer profile of various taxi: K-means clustering, DBSCAN, Hierarchical clustering, STING. To investigate the impact of weather and time on driver's daily income, we will implement hypothesis testing.

References

1. Todd W. Schneider. November 2015. Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. Retrieved from: https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

2. Kaggle. 2017. New York City Taxi Trip Duration.

3. Shashank Badre. December 03, 2017. Exploratory data analysis on Green Taxi. Retrieved from:http://rstudio-pubs-static.s3.amazonaws.com/326454_0e4d6355b75a4578bebac6cd99cc319f.html

4. Chih-Ling Hsu. May 14, 2018. Analyze the NYC Taxi Data. Retrieved from: https://chih-ling-hsu.github.io/2018/05/14/NYC#q1-which-regions-have-most-pickups-and-drop-offs

5. Willy Sebastian. June 2, 2018. New York Taxi Trip Analysis. Retrieved from: https://rpubs.com/willyarrows/NYCTaxiTripsAnalysis

[2] Yellow Trips Data Dictionary: https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf